4-2012

Taxonomic Classification of Media Reports in the Cyber Attack Domain

Jinhua Zhang
University of Nebraska at Omaha

Taxonomic Classification of Media Reports in the Cyber Attack Domain

A Thesis Presented to the

Department of Computer Science and the Faculty of the Graduate College

University of Nebraska

In Partial Fulfillment of the Requirements for the Degree

Masters of Science

University of Nebraska at Omaha

By

Jinhua Zhang

April, 2012

Supervisory Committee

Qiuming Zhu

William Mahoney

Robin A. Gandhi

Stanley Wileman

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Taxonomic Classification of Media Reports in the Cyber Attack Domain

Jinhua Zhang, Master of Science

Graduate College, University of Nebraska, 2012

Advisor: Qiuming Zhu

Abstract

Cyber-attacks have become a huge threat to the information age. In a previous study, cyber-attacks associated with events in Social, Political, Economic and Cultural (SPEC) dimensions were analyzed [6]. The task of this research is to construct an automated classifier that can classify media reports related to past and current cyber-attack events according to the SPEC taxonomy. The classifier was built on a machine learning principle incorporated with approaches focused on 1) document indexing; 2) calculation of classification thresholds; 3) definition of classification effectiveness; and 4) calculation of precision and recall. The classifier is expected to perform with acceptable effectiveness so that it helps analysts to deal with large number of media reports that are predictive of possible cyber-attacks.

Acknowledgments

This thesis owns a lot to the inspiration, encouragement, suggestions and criticism of my professors Dr. Qiuming Zhu, Dr. William Mahoney, Dr. Robin A. Gandhi and Dr. Stanley Wileman. I also want to thank my fiancée Wenyu Shi for her deepest love and unwavering support throughout my study and life.

Contents

List of Figures

List of Tables

1. Introduction and Related Work

Our daily life is increasingly computer- and Internet-based: Computers and their networks now permeate industries and business services such as commodity production, transportation, healthcare, finance, infrastructure management etc. While computing technologies have made information access, processing and exchange easier and more efficient, however, it has also created a new space in which criminals and terrorists can operate almost undetected. Cyber security has becomes crucial for the social, political, economic and cultural safety of a nation. As President Obama has stressed [3], the ever-growing number of attacks on cyber networks has become "one of the most serious economic and national security threats our nation faces."

A large number of researches and practitioners are taking various efforts attempting to prevent cyber-attacks. As a currently prevalent prevention mechanism, anomaly detection models focus primarily on analyzing network traffic to prevent malicious activities [5, 10, 11, 15]. However, one important factor is overlooked: A cyber-attack is a highly complicated sociotechnical event behind which the humans operating it are actually the most decisive component, for whose most behavior anomaly detection models fail to account. And these humans, known as cyber attackers, are often associated with social, political, economic, and cultural (SPEC) conflicts [14, 18, 19]. It has been increasingly realized that by investigating cyber-attackers' level of socio-technological sophistication, their backgrounds, and their motivations in those

four dimensions, cyber-attacks can be more effectively predicted, prevented, and traced [12, 16].

Researchers from University of Nebraska at Omaha have analyzed potential correlations between historical and current cyber-attacks in relation with the Social, Political, Economic, and Cultural (SPEC) events, and provided valuable insights regarding to the origin, agents, means, motives, and potential targets of future cyber-attacks [7]. In their previous study, they described how several SPEC events have led to cyber-attacks, and then presented a taxonomy and analysis of these attacks in the SPEC dimensions.

However, due to the tremendous amount of text information regarding cyber-attacks available not only in the past media reports but also published continuously almost every day, it is an incredibly huge workload to categorize this information according to the SPEC taxonomy. To do this only by human analysis would be, if not impossible, extremely arduous and time-consuming. On the other hand, similar tasks can be tackled with much more ease and efficiency by utilizing techniques of machine-learning-based automated text classification. There is already a large body of related work and various methods have been proposed in the past [1, 3, 4, 6, 13, 17].

This study represents another attempt of the Machine Learning Approach, which has gained popularity since the early 90's and has eventually become the dominant method today [13, 17]. The classification task in this study is a multi-label one [17], which means any number (>= 0) of categories can be assigned to one single document, since one cyber-attack can have more than one of the SPEC dimensions. Because of

the targeted use for the tool, the classifier to be built in this study is a document-pivoted one [17], which try to find all the applicable categories from the four SPEC categories. This is contrasted with a category-pivoted approach where the attempt is to locate documents for one given category. The reason for the adoption of the former way is that in this very operation context, the four dimensions of cyber-attacks are relatively fixed; while cyber-attack records are updated quite frequently as new cyber-attack events continuously take place.

The paper is organized as follows: In Section 2, we introduce the concept of the initial text corpus; in Section 3 we describe the learning part including our key-word list document indexing approach, term weight calculation, as well as threshold calculation. Two term weight calculation methods were proposed and their results were compared. In the end of section 3, we give the pseudo codes to implement the classifier in terms of categorization of new-coming cyber-attack events. The classifier built is evaluated in Section 4, where we describe our six matching types of the classification result and the expert decision as well as our innovative calculation methods of precision and recall. Section 5 gives our three experimental tests and their results. We give a brief overview of the user interface of the classifier in form of a web application in Section 6. At last, we summarize the whole study and discuss some future work directions in Section 7.

2. The initial corpus and the manual pre-classification

The initial text corpus of this study contains 45 recent (2000-2011), worldwide cyber-attack records, which are typical and influential along one or more of the SPEC dimensions. The resource of those records includes newspapers, periodicals, TV, Internet and so on. Through careful investigation and selection, their truthfulness is guaranteed, while their most important details including attack time, origin location(s), target location(s), affected location(s), event description, attacker(s), victim(s) and casualty are available for the most cases. These records are manually pre-classified into the SPEC categories based on expert knowledge, namely the Social, Political, Economic, and Cultural characteristic identifications, with some events overlapping across multiple categories. The final look of the pre-classified-corpus is partly shown in Figure 2-1 below. Complete details of this corpus will be given in section 5.

Figure 2-1 The Pre-Classified Text Corpus

Recent Cyberattacks | Cyberattack Events | Advanced Search | Add New Events | Metrics Page

Description	Type	Date	SourceLocation	TargetLocation	Attacker	Victim	Casualty
HBGary's systems and website hacked by Anonymous	E	Feb 7, 2011		USA	Anonymous	HBGary company	HBGary's tech support servers was broken then the company was defrauded of lots private informatioin by hackers
WikiLeaks Says Was Denial-Of-Service Attack Victim	S/P	Nov 24, 2010	n/a	USA	Unknown Hackers	WikiLeaks	hackers trying to prevent its release of classified U.S. State Department documents
Human rights organization targeted with cyber attack	S/C	Nov 3, 2010	Indonesia	n/a	Indonesian	The website of Survival International	
Was China behind cyber attack on Nobel Peace Prize website?	S/P	Oct 27, 2010	n/a	Norway	N/A	Nobel Institute	Computers users receiving emails
WikiLeaks Releases 390,000 Iraq War Documents	S/P	Oct 25, 2010	USA	USA	wikileaks	American government	391,832 secret Iraq War documents were published
Judge Clears CAPTCHA-Breaking Case for Criminal Trial	E	Oct 19, 2010	USA	USA	American Hackers	Related business websites	Disrupt the normal market order
School District Pays $610,000 to Settle Webcam Spying Lawsuits	S/E	Oct 12, 2010	USA	USA		students in Philadelphia school	lots of private pictures lost due to webcam spying
Hacked Voting System Stored Accessible Password, Encryption Key	S/P	Oct 6, 2010	USA	USA	researchers at the University of Michigan	internet voting system of Washington,DC	database username, password and encryption key were lost
Secret-Spilling Sources at Risk Following Cryptome Breach	E	Oct 5, 2010	USA	USA	the group Kryogeniks	Secret-spilling site Cryptome	lost lots of secret files and correspondence

3. Learning and Classifying

The learning phase consists of several components that link together which include document indexing, key-word list establishment and threshold determination. We will also present the pseudo codes for the classifying phrase.

3.1 Document Indexing

An indexing procedure maps a document into a compact representation of its content [17]. In this study, *indexing terms* are identified with individual words rather than phrases. Before indexing, the *stop words* were removed according to the stop-word-list sorted out by MIT (571 words) [8], and stemming was performed by the Porter's algorithm [9] in advance. An *event* described in one particular document is represented by one *bag of words*, such that

$$event_i = \{k_{i1}, k_{i2}, \ldots k_{ij}\},$$

where k_{ij} is the j^{th} stemmed word of the i^{th} text of the corresponding corpus.

The next step is to establish a key-word list, which is a two-dimensioned table composed of 1)words from the *bags of words* but with word repetition eliminated, called *key words*, 2) *term weights* of these key words with regard to SPEC taxonomy. More specifically speaking, each keyword in the keyword list has four properties: S(Social), P(Political), E(Economic) and C(Cultural), corresponding to the four cyber-attack categories. For each property, a *term weight* is assigned, which represents the degree of contribution of a given keyword to the semantics of that

category. Table 3-1 shows one way to represent a key-word list, where wk_{1s} is the term weight of key word key_1 to the property (category) of S.

Table 3-1 Key-word List

Key Word	S (Social)	P (Political)	E (Economic)	C (Cultural)
key_1	wk_{1s}	wk_{1p}	wk_{1e}	wk_{1c}
key_2	wk_{2s}	wk_{2p}	wk_{2e}	wk_{2c}
...				
key_n	wk_{ns}	wk_{np}	wk_{ne}	wk_{nc}

In 3.1.1, we will elaborate on the establishment procedures of the key-word list.

3.1.1 Key-word List Establishment Procedures

The key-word list is established through the procedures as described in the following steps.

Step 1: Forming integrated keyword list Set_{temp}

As there are *n* documents in the training set, *n* bags of words are created, one for each document. Note that each document describes one event of interest. After the bag of words for each document is created, an integrated word list, named Set_{temp}, is formed by putting all the words from the documents together, such that.

$$\begin{aligned} Set_{temp} &= \{event_1, event_2 \ldots event_n\} \\ &= \{\{k_{11}, k_{12}, \ldots k_{1x}\}, \{k_{21}, k_{22}, \ldots k_{2y}\}, \ldots \{k_{n1}, k_{n2}, \ldots k_{nz}\}\} \\ &= \{k_{11}, k_{12}, \ldots k_{1x}, k_{21}, k_{22} \ldots k_{2y}, \ldots k_{n1}, k_{n2}, \ldots k_{nz}\} \end{aligned}$$

Where $\{k_{i1}, k_{i2} \ldots k_{ix}\}$, i = 1… n, represents a bag of words from document i, i.e., the event described in document i.

Step 2: Forming word duplication removed keyword list *ks*

If there is a word k_{nz} appearing more than once in the Set_{temp}, the repeated

word is eliminated from the list. The resulting *keyword set*, denoted as *ks*, is obtained:

ks= {key1, key2 ...keym}, where key_m is the unique appearance of k_{nz}.

Step 3: Computing term weight wk_{1s},

In this step, we will figure out the term weight of key_m regarding to each category (S/P/E/C) by working out its TF-IDF (Term Frequency–Inverse Document Frequency) weight [16] to that category. Now we are taking wk_{1s} (See Table 3.1) as an example.

In order to obtain wk_{1s}, namely the weight of the first keyword (key_1) of the *keyword set* (*ks*) for the category of "Social", two methods are experimented.

3.1.1.1 Term Weight Computation Method 1

Step 1): Based on the expert decision, if there are *p* bags of words under the category "Social" of the initial corpus, represented as $event_1, event_2, \ldots event_p$, we first merge these *p* bags of words together as one bag of words Event_S

Step 2): Then we calculate the wk_{1s} by:

$$wk_{1s} = tf \times idf\ (t, d, D) = \frac{t}{d} \times \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

where t is the number of occurrences of key_1 in bag of word Event_S;

d is the word count in the bag of words Event_S;

D is the category count of the training set;

$|\{d \in D : t \in d\}|$ is the number of category that contains key_1.

3.1.1.2 Term Weight Computation Method 2:

Step 1): Based on the decision of the expert decision, if there are *p* bags of words

under the category "Social" of the initial corpus, represented as $event_1$, $event_2$, ... $event_p$, we apply the TF-IDF computation to get the weight of key_1 with respect to $event_1$, such that

$$wk_{1s}^{event_1} = tf \times idf(t, d, D) = \frac{t}{d} \times \log \frac{|D|}{|\{d \in D : t \in d\}|},$$

where t is the number of occurrences of key_1 in $event_1$;

d is the word count of $event_1$;

D is the event count of the training set;

$\{d \in D : t \in d\}$ is the number of events in the training set that contains key_1.

Step 2): Calculate the overall weight by

$$wk_{1s} = \frac{wk_{1s}^{event_1} + wk_{1s}^{event_2} + \ldots wk_{1s}^{event_p}}{p}$$

Following the same steps, we obtained the weight value of each keyword (key_m) to each category (S, P, E, and C) until it is done for the entire keyword list.

3.1.1.3 Differences of the two methods

The major differences of the above two methods are described: Instead of calculating the TF-IDF weight of each key word regarding to each event and averaging the sum of weight of all the events, as it does in method 2, method 1 merges all of the events under the same category into one event. Since one event exists in the form of one *bag of words*, this means we merge all the bags of words of the same category, let's say, Social, into one general bag of words of the category Social. By doing so the training set is represented by four large bags of words: Event_S, Event_P, Event_E and

Event_C. Our experiments show that the results of Method 1 are less accurate than Method 2, thus Method 1 is not as desirable for the classification task we are attempting. This can also be seen from the following analysis: Although the composition of these bags of words after all remains the same, the manner in which the words are collected into the bags influences the weight computation. When we apply Method 1 to calculate the term weight of each key word to each category by:

$$wk_{1s} = tf \times idf\ (t, d, D) = \frac{t}{d} \times \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

where t is the number of occurrences of key_1 in Event_S;

d is the word count of Event_S;

$|D|$ is the category count of the corpus;

$|\{d \in D : t \in d\}|$ is the number of category that contains key_1.

Apparently, the value of $|\{d \in D : t \in d\}|$ has only four possibilities: 1, 2, 3, and 4, and the category count of the corpus $|D|$ is 4, which makes the computational results highly biased. What's worse, as the size of the training set grows; the possibility of one given key word of appearing in all the four categories grows at the same time. This causes the value of $|\{d \in D : t \in d\}|$ which is 4 to grow, resulting in plenty of 0 values in the term weight as shown in the following equation:

$$wk_{1s} = tf \times idf\ (t, d, D) = \frac{t}{d} \times \log \frac{|D|}{|\{d \in D : t \in d\}|} = \frac{t}{d} \times \log \frac{4}{4} = 0$$

In another words, many key words become useless and thus reduce the efficiency of the classifier. Test 1 will demonstrate this inferiority and undesirability.

3.2 Classification with Thresholds

3.2.1 The Principle of Threshold Determination

Threshold is a cut-off value that the classifier uses to decide to which category (or categories) a document should be assigned. For the four categories of Social, Political, Economic, and Cultural, there are four corresponding thresholds, represented as threshold_s, threshold_p, threshold_e and threshold_c. When a new event is processed, the classifier will first work out a value that represents the possibility of this event of belonging to one particular category, named as "*resemblance*". Then we use four values that represent the *resemblance* of this event with respect to the four categories respectively, denoted by r_s, r_p, r_e and r_c respectively. The classifier compares, for instance, r_s with threshold_s, if r_s > threshold_s, the classifier will attribute this event to Category S. In the learning phase, we already know the expert decision, i.e. in which category (or categories) an event should be assigned, and we will obtain r_s, r_p, r_e and r_c later on. Based on those two values, we infer in a step-back manner in order to determine the threshold.

In the following sub-section, we will elaborate on this process.

3.2.2 Threshold Determination

Let's look at how to determine threshold_s as an example.

After getting the key word list, a variable r_1_s is assigned to the first word for the first bag of words ($event_1$). Then we look for its term weight in the key-word list and add this weight value to the variable r_1_s. While traversing the rest of the word in $event_1$,

we continue to sum up and get the final value of r_1_s, which represents, as mentioned previously, the possibility of $event_1$ belonging to Category S.

In Table 3-2, r_1_s represents the possibility of *event*$_1$ belonging to Category S. Similarly, we obtain the resemblance for each event in terms of each category.

Table 3-2 the Resemblance Table

Event	S(Social)	P(Political)	E(Economic)	C(Cultural)
$event_1$	r_1_s	r_1_p	r_1_e	r_1_c
$event_2$	r_2_s	r_2_p	r_2_e	r_2_c
...				
$event_i$	r_i_s	r_i_p	r_i_e	r_i_c

The resemblance of all the events in terms of Category S will be compared to the manual pre-classification results, also called *expert decisions*, one by one. Among the *positive examples* of the training set in terms of Category S, i.e. the events that according to the expert decisions belong to Category S, the smallest one of their resemblance values to S, denoted as $r\_s_{smallest}$, will be drawn out. Among all the other events, i.e. the negative examples of the training set, the largest one of their resemblance values to S, denoted as $r\_not\_s_{largest}$, will be drawn out. We believe that the threshold_s exists between the two. For easy of determination, we decide on the arithmetic mean of the two, i.e.:

$$threshold\_s = \frac{r\_s_{smallest} + r\_not\_s_{biggest}}{2}$$

Theoretically, $r\_s_{smallest}$ is always larger than $r\_not\_s_{largest}$, however, in practice, situations do exist where $r\_s_{smallest} <= r\_not\_s_{largest.}$ This is because there is an error between the results of text classification and the manual (pre-) classification; there are

two reasons for this:

1) The TF-IDF algorithm itself has a certain degree of error;

2) Manual classification also has a certain degree of error. For example, when an expert believes that one event belongs to Category S, it is only a subjective judgment, which is not necessarily 100% true. Often times a certain category will just seem more "reasonable" to the expert. This effect tends to increase when one event can be attributed to more than one category, say, S and P. To the contrary, as the algorithm performs the classification, as long as the needed mathematical formula is satisfied, it can determine that an event belongs to S with 100% certainty, and also belongs to P with 100% certainty.

If the discussed error reaches a considerable, influential level so that a situation where $r\_s_{smallest} <= r\_not\_s_{largest}$ does happen, the following measure should be taken: Among the *positive examples* of the training set in terms of Category S, the *second smallest* of their resemblance values to S will be drawn out. Among the negative examples of the training set, the *second largest* of their resemblance values to S will be drawn out. If the former is still the same as or smaller than the latter, the corresponding pair in the third, fourth, fifth… place will be tried out, until the required inequality is satisfied. In this case, if we can counteract to the error with the p^{th} pair, we say that the *error rate* of *threshold_s* is $\frac{p}{n}$, where *n* is the event count of Category S in the initial corpus.

3.3 Classification

The process of classifying can be implemented through the following pseudo code:

```
//input a document
document d;
// get the bag of word from the given document
event_new = index(d) = {k1,k2...kn};
// initialize the resemblance value for each category
r_s, r_p, r_e, r_c=0;
// get the key word set from the key word list
ks={key1,key2...keym};
// get the threshold of each category
threshold_s,threshold_p,threshold_e,threshold_c;
// initialize the result;
type = null;
// for each word in event_new, find if it exists in key word set, if so,
// add the term weight of each category to the corresponding resemblance.
for (int i=0;i<n;i++) {
    for(int j=0;j<m;j++) {
        if(ki = keyj) {
            r_s += wkjs;
            r_p += wkjp;
            r_e += wkje;
            r_c += wkjc;
        }
    }
    r_s = r_s/n;
    r_p = r_p/n;
    r_e = r_e/n;
    r_c = r_c/n;
// compare the resemblance value of each category with its corresponding
// threshold, and add the result to type.
    if(r_s >threshold_s)
    type = S;
    if(r_p >threshold_p)
    type = type + P;
    if(r_e >threshold_e)
    type = type + E;
    if(r_c >threshold_c)
    type = type + C;
// return the result
return type;
}
```

4 Evaluation

In this section we evaluate the classifier built by the machine learning process described above. We perform three experimental tests for this purpose: Test 1 is aimed to demonstrate the undesirability of Learning Method 1 elaborated in 3.1.3, and demonstrate the effectiveness of Method 2; Test 2 is aimed to measure the

performance of the classifier based on Method 2 in terms of the classification effectiveness as the training set grows; Test 3 is aimed to further test the effectiveness of the classifier based on Method 2 with respect to events randomly drawn from public media resources.

4.1 The Criteria in terms of Effectiveness – Matching Types

The Effectiveness of a classifier is defined as the ability to take the *correct* classification decisions [16]. In this study, a six-degree scale was created to define the "*correctness*" by measuring how much the classification results, or text classification (TC for short) match the expert decision (ED for short). Listed from the highest to the lowest degree of matching, the six *matching types* are (also as shown in Table 4-1):

a. **Exact Match:**

   An exact match indicates that the auto-classification result is identical to the expert decision.

   E.g. TC: Social, Political, Economic, while ED: Social, Political, and Economic;

b. **Membership Match:**

   A membership match means that the classification result forms a subset of the expert decision.

   E.g. TC: Social, Political, while ED: Social, Political, Economic;

c. **Intersection Match:**

   An intersection match occurs when the classification result and the expert

decision share a certain number of elements (categories).

E.g. TC: Social, Political, while ED: Social, Economic;

d. Orientation Match:

An orientation match occurs when the classifier returns NULL as result, but the distribution of the resemblance degrees of the event in terms of the four categories shows an orientation that echoes the expert decision.

E.g. TC: NULL, with $r_i_s<$threshold_s, $r_i_p<$threshold_p, $r_i_e<$threshold_e, $r_i_c<$threshold_c, while $r_i\_s>r_i\_e$, $r_i\_s>r_i\_c$, $r_i\_p>r_i\_e$, $r_i\_p>r_i\_c$; ED: Social, Political;

e. Distinction Match:

A distinction match means that there is no overlap between the category (or categories) derived from the classifier and the category (or categories) determined by the expert decision.

E.g. TC Result: Economic, Cultural, while ED: Social, Political;

f. No Match:

Finally, no match implies that the classifier returns NULL as its result, while the distribution of the resemblance degrees of the event in terms of the four categories disagrees with the expert decision.

E.g. TC: NULL, with $r_i_s<$threshold_s, $r_i_p<$threshold_p, $r_i_e<$threshold_e, $r_i_c<$threshold_c, while $r_i\_s<r_i\_p$; ED: S.

Table 4-1 Six Matching Types

Match type	Description
1. Exact Match	The result of the classifier (short for TC) is identical to the expert decision(short for ED).
2. Membership Match	TC forms a subset of ED.
3. Intersection Match	TC and ED shares a certain number of elements.
4. Orientation Match	TC is NULL, but the distribution of the resemblance degrees of the event in terms of the four categories shows an orientation that echos the ED.
5. Distinction Match	There is no overlap between TC and ED
6. No Match	TC is NULL, while the distribution of the resemblance degrees of the event in terms of the four categories disagrees with ED.

Among the six situations described above, if a classifier reaches a result of 1) Exact Match, or 2) Membership Match, or 3) Intersection Match, or 4) Orientation Match with a certain event, we believe that the classifier makes a correct decision on that event. The Orientation Match is kept because, as its name suggests, it can offer an orientation, a reference for human classification that might be of more or less help especially in the case where a corpus is in its early stage of development.

All the classification results that represent the classification achieves in terms of Exact Match, Membership Match, Intersection Match, Orientation Match, Distinction Match and No Match will be assigned into six sets, named as S_e, S_m, S_i, S_o, S_d, and S_n respectively. The classifier is tested using a labeled test set. That is, the categorizations of the documents are assumed to be known before the test, these are denoted as the number of ED results in Set_{ed}. In the test, the number of results in each of the six matching types is accumulated.

4.2 Measures of Effectiveness

Classification effectiveness is usually measured in precision (π) and recall (ρ). Precision with regard to a certain category $C_x(\pi)$ is defined as the probability that a random document d_i is correctly classified under C_x. Analogously, recall with regard to $C_x(\rho)$ is defined as the probability that a random document d_i should be classified under C_x and this decision is taken. In other words, π can be viewed as the "degree of soundness" of the classifier with regard to C, while ρ may be viewed as its "degree of complete" with regard to C [17].

Conventionally, *precision (π)* and *recall (ρ)* can be estimated by means of the conventional model of "*contingency table wrt C_i*", in which concepts of true positives wrt C_i (TP_i), false positives wrt *Ci*(*FPi*), true negatives wrt C_i (TN_i) and false negatives wrt C_i (FN_i) are included (see Table 4-2) [17].

$$\hat{\pi}_i = \frac{TP_i}{TP_i + FP_i} \qquad \hat{\rho}_i = \frac{TP_o}{TP_i + FN_i}$$

where TP_i is the number of the documents correctly classified into C_i,

FPi is the number of the documents incorrectly excluded out of C_i,

FN_i is the number of the documents incorrectly classified into C_i.

(TN_i is the number of the documents correctly *excluded out of C_i*.)

So the overall precision and recall of the classifier are:

$$pricision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

Table 4-2 Contingency Table for Category Ci [17]

Category c_i		expert decision	
		YES	NO
Classifier judgments	YES	TP_i	FP_i
	NO	FN_i	TN_i

However, this model was not adopted in this study because the precision and recall obtained this way is related to only one category C_i. Further, using this method, the classification result for one particular category C_i with respect to one event has only two possibilities of being either “correct” or “incorrect”. To be “correct” means that the classification result must be an exactly 100% match to the expert decision, while “incorrect” means that the classification result doesn’t agree with the expert decision at all, that the match degree is 0%.

Apparently, the conventional precision and recall model cannot be directly used for the situation of our approach: Here, the classification result with regard to one event involves four categories rather than only one. It is no longer a simple issue of “either correct or incorrect”. For example, an event is attributed by the classifier to Category S, which is correct, but it has nothing to do with whether it also belongs to Category P. But by “correct”, our expectation is that the classifier can find all the categories applicable to the event. That’s why we created the six matching types and confined “correct” to the first four highest match degrees in 4.1. In our system, a 100% match of the classification result and the expert decision is called an “Exact Match”, and a 0% match of the two is equal to our “Distinction Match” or “No Match”. Now, our task is to assign a calculable percentage to the other three situations, namely Membership Match, Intersection Match and Orientation Match. For the first two, we

simply compare the classification result and the expert decision, both with regard to one event.

For example, for $event_1$, we have a situation as shown in Table 4-3.

Table 4-3 Comparison between Classification Result and Expert Decision

Categories	S	P	E	C
TC	S	P	Null	Null
ED	S	Null	E	Null
Match	Same	Not same	Not same	Same

So the *matching degree wrt event*$_1$ is $\frac{1+1}{4} = 0.5$

Generally speaking, we need to figure out the *matching degree wrt event*$_t$, denoted as $md(event_t)$, then our precision measurement is defined as:

$$precision = \frac{md(event_1) + md(event_2) + \ldots + md(event_t) + 0.5 \times o}{ed}$$

$$where\ event_t \in \{S_e, S_m, S_i\}$$

$0.5 \times o$ means that for each event which achieves an Orientation Match, we always assign a value of 0.5 to that event as its matching degree.

As for the calculation of *recall*, the reason that we did not adopt the conventional method is the same reason that we did not do it with *precision*. Again, even though a result returned by Orientation Match does not indicate any particular category, given its referential meaning for human classification, we regard it as a correct algorithm decision. Our recall measurement thus is defined as:

$$recall = \frac{e + m + i + o}{e + m + i + o + d}$$

where e is the count of classification results that achieve Exact Match,

m is the count of classification results that achieve Membership Match,

i is the count of classification results that achieve Intersection Match,

o is the count of classification results that achieve Orientation Match,

d is the count of classification results that achieve Distinction Match.

5 Effectiveness Tests

In this section we present the results of our three test cases using the precision and recall measurements described in the last section.

Details of the initial text corpus that we use in the test are shown in Table 5-1.

Table 5-1 Initial Text Corpus

Case #	Event title	Category
1	Japanese textbook dispute sparks cyber attack	SPC
2	Hackers Stole IDs for Attacks	SP
3	French embassy in Beijing under cyber-attack after Nicolas meeting with Dalai Lama	SPC
4	China analysts dismiss cyber espionage claims	PE
5	Cyber attackers empty business accounts in minutes	E
6	US websites buckle under sustained DDoS attacks	P
7	FBI to investigate Placentia Library hacking	P
8	New Virus Appears As Response To Craigslist Ad	S
9	Targeted Malware Attack on Foreign Correspondents based in China	S
10	Polish government cyberattack blamed on Russia	S
11	Attack Hits Swedish Signals Intelligence Agency's Website	S
12	Cyber vandal hits police website	P
13	Climate Change E-mail Hack Could Lead To Future Attacks	SP
14	Baidu hacked by Iranian cyber army	SP
15	Chinese human rights Web sites suffer attacks	SP
16	Government sites crumple under Operation Titstorm's DDoS attack	S
17	Two Koreas in Cyber Proxy War	P
18	Hacker defaces Iowa Homeland Security web site forces shutdown	S
19	Cyber attack shut 150 Montenegrin websites	P
20	Westin Hotel's POS Hacked	E
21	St. Louis police department hit by cyber attack	SE
22	Cyber attack brought down national election website	S
23	Google Links Web Attacks to Vietnam Mine Dispute	P
24	Web site of China-based journalist club attacked	SP
25	Researchers Trace Data Theft to Intruders in China	P
26	Cyber criminals quick to pounce on McAfee crash story	E
27	1.5M stolen Facebook IDs up for sale	E
28	Fake fast food survey with cash reward leads to phishing site	E
29	Koobface server pops up in China after HK takedown	SE
30	Cyber attack lands G.I. man in jail	C
31	Hackers shut down EU carbon-trading website	S
32	Defaced gov't websites another black eye for RP	SP
33	Burned in Sex Sting, Hacker Attacks Computers	SC
34	Burmese websites attacked by hackers	S

35	5 Key Players Nabbed in Ukraine in $70-Million Bank Fraud Ring	S
36	Intel chief says Iran able to fight off worm that hit computers linked to nuclear plant	SPC
37	Secret-Spilling Sources at Risk Following Cryptome Breach	E
38	Hacked Voting System Stored Accessible Password, Encryption Key	SP
39	School District Pays $610,000 to Settle Webcam Spying Lawsuits	SE
40	Judge Clears CAPTCHA-Breaking Case for Criminal Trial	E
41	Biggest Military Leak in History: WikiLeaks Releases 390,000 Iraq War Documents	SP
42	Was China behind cyber attack on Nobel Peace Prize website?	SP
43	Human rights organization targeted with cyber attack	SC
44	WikiLeaks says was denial-of-service attack victim	SP
45	HBGary's systems and website hacked by Anonymous	E

In this corpus, there are 28 records of the Social category, 24 of the Political category, 13 of the Economic category and 5 of the Cultural category.

5.1 Test 1

As mentioned at the beginning of the previous section, Test 1 is aimed to demonstrate the inferiority and undesirability of Learning Method 1 elaborated in 3.1.2, compared to Method 2 described in the same section.

Prior to classifier construction, an initial text corpus with 45 manually pre-classified cyber-attacks events was split in to two sets - one training set and one test set. There are 35 randomly selected events in the training set, while the rest 15 formed the test set. The classifier was built by the two learning methods respectively and their classification results were compared. This process was repeated 10 times.

The results are shown in the following table (Table 5-2):

Table 5-2 Comparison between the results of Method 2 and Method 1

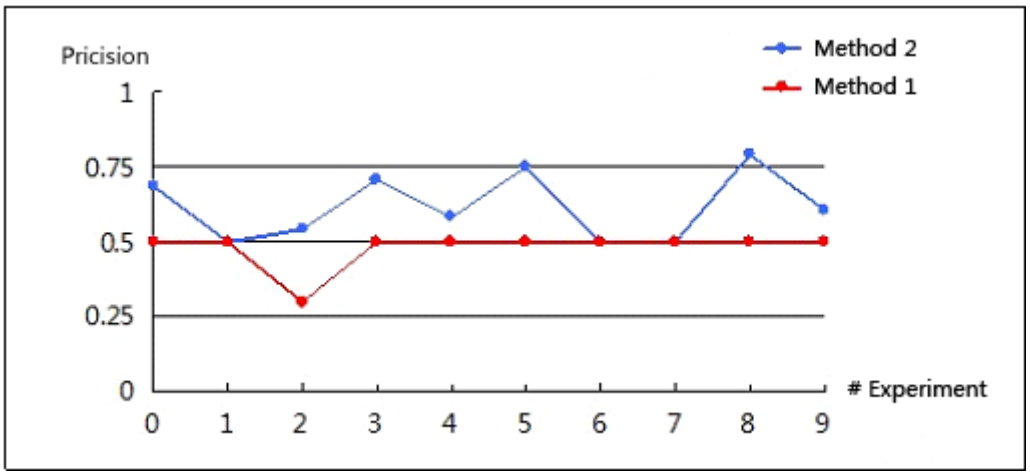
#Experiment	Method 2		Method 1	
	Weighted Precision	Recall	Weighted Precision	Recall
1	0.688	0.4	0.500	0.4
2	0.500	0.4	0.500	0.4
3	0.542	0.6	0.300	0.5

4	0.708	0.6	0.500	0.6
5	0.583	0.6	0.500	0.6
6	0.750	0.7	0.500	0.5
7	0.500	0.7	0.500	0.3
8	0.500	0.4	0.500	0.2
9	0.792	0.6	0.500	0.2
10	0.607	0.7	0.500	0.4

This table can is transformed into Line Chart 1 and 2 for easier comparison:

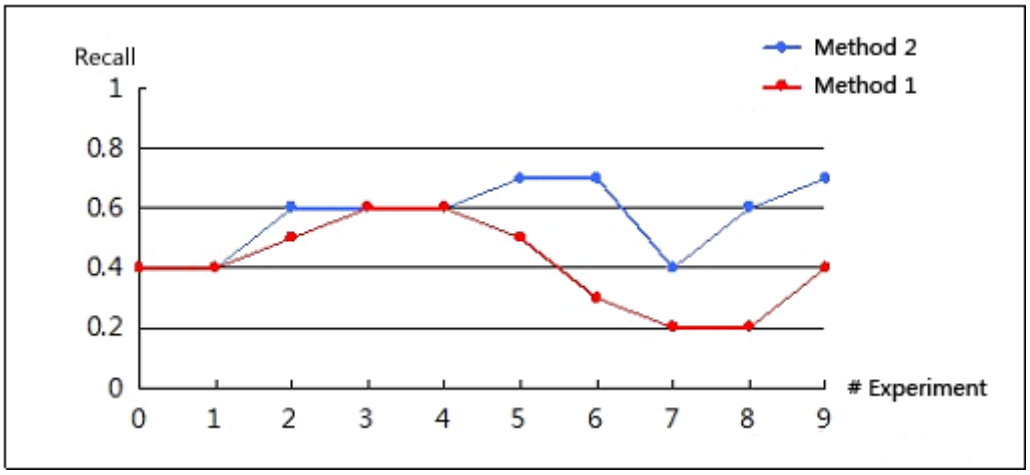
1) Precision

Figure 5-1 Comparison the Precisions of the Two Learning Methods



2) Recall

Figure 5-2 Comparison the Recalls of the Two Learning Methods



5.2 Test 2

Test 2 is aimed to observe the change in the classification effectiveness as the training set grows. The initial text corpus was split into four sets, Set_1 .. Set_4, in which there are respectively 30, 5, 5 and 5 randomly selected events. The test performed the following steps:

1) The classifier was built by learning Set1 as training set, and was tested on Set4 as test set.

2) The classifier was rebuilt by learning the merged set of Set1 and Set2 as training set, i.e., the training set size was increased. The training result was tested on Set4 as test set.

3) The classifier was rebuilt again by learning the merged set of Set1, Set2 and Set3 as training set, i.e., the size of the training set was increased. The training result was tested on Set4 as test set.

4) The same procedures above were repeated 10 times.

The results are shown in the following table (see Table 5-3):

Table 5-3 Classification Result of Test 2

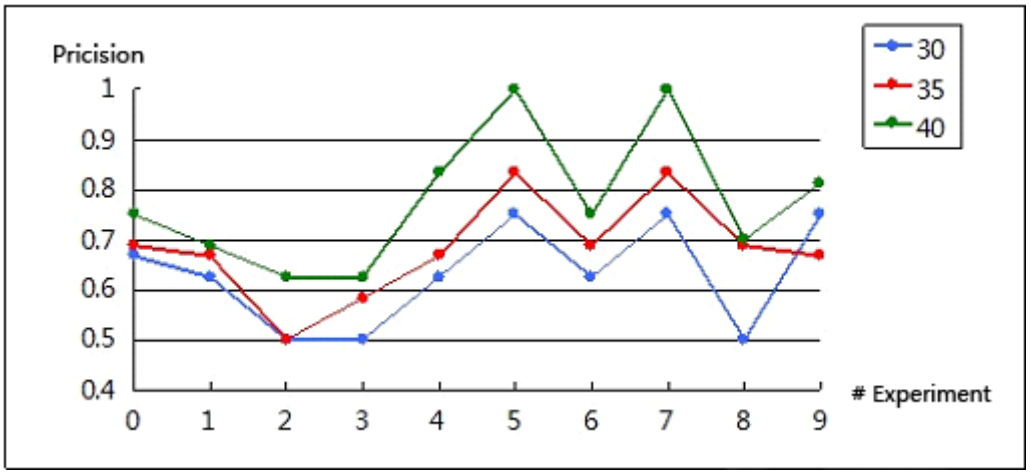
#Experiment	Size of Training Set: 30 events		Size of Training Set: 35 events		Size of Training Set: 40 events	
	Precision	Recall	Precision	Recall	Precision	Recall
0	0.667	0.6	0.688	0.8	0.750	0.8
1	0.625	0.4	0.667	0.6	0.688	0.8
2	0.500	0.2	0.500	0.2	0.625	0.4
3	0.500	0.4	0.583	0.6	0.625	0.8
4	0.625	0.4	0.667	0.6	0.833	0.6
5	0.750	0.6	0.833	0.6	1.000	0.8

6	0.625	0.6	0.688	0.8	0.750	0.8
7	0.750	0.4	0.833	0.4	1.000	0.6
8	0.500	0.8	0.688	0.8	0.700	1.0
9	0.750	0.4	0.667	0.6	0.813	0.8

This table is transformed into Line Chart 3 and 4 for easier comparison:

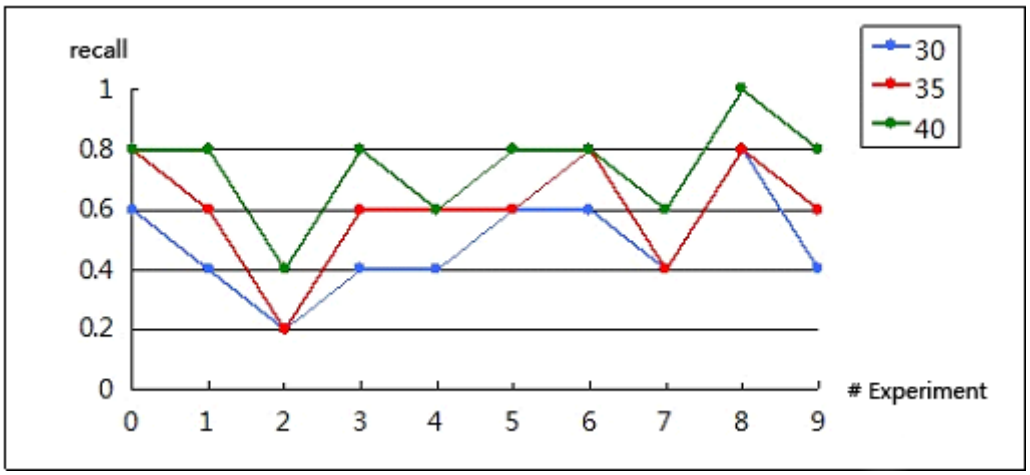
1) Precision

Figure 5-3 Precision of Test 2



2) Recall

Figure 5-4 Recall of Test 2



Conclusion of Test2: As the training set grows, the effectiveness with the proposed learning method also grows.

5.3 Test 3

Test 3 is aimed to recognize the general effectiveness of the classifier to randomly selected events from public data resources (outside of the initial corpus). The initial corpus with 45 manually pre-classified events served as the training test. Another corpus with 100 new events was created. Some of these events can be classified into one or more of the SPEC categories, while a few were deemed to not belong to any. The set was initially manually pre-classified based on expert decision, and then automatically classified by the classifier constructed on the training set. The achieved overall weighted precision was 63.94% and the recall was 66.25%. Table 5-4 shows the details of the 100 new events, while Table 5-5 shows the result of the classification.

Table 5-4 Details of the 100 New Events

Case #	Event title	Category		Case #	Event title	Category
The first 80 events are used to test for relevance to SPEC				51	Medic Hackers Stole Confidential Data	E
				52	Military contractors targeted in Chinese attacks says F-Secure	PE
1	Analysis Who cyber smacked Estonia	P		53	More than 100 companies targeted by Google hackers	PE
2	CIA Admits Cyber attacks Blacked Out Cities	SP		54	N Korea Suspected of Global Cyber Attack	P
3	Climate Change E-mail Hack Could Lead To Future Attacks	S		55	NATO boosts cyber-attack response force senior official	SP
4	CNN site hit by China attack	SP		56	China blames US cyber attack for Iran unrest_scots man	PC
5	Coming to terms with cyber warfare	P		57	Olympics-Cyber attack seen as emerging threat for London	SPE
6	Could US repel a cyber attack	SP		58	Other countries developing cyber attack capability CIA says	P
7	Crackdown on Mariposa Botnet Infected 13 Million PCs	E		59	OtherCountriesDevelopingCyberAttackCapabilityCIA Says	P
8	Cracks in the System	SC		60	Over 75000 systems compromised in cyber attack	SE
9	Cyber Attack Data-Sharing Is Lacking Congress	SP		61	Pentagon Bill To Fix Cyber Attacks 100M	P
10	Cyber attacks target Japan sites	SPE		62	Political Cyber Attacks Hit Half Of Large Companies	PE
11	Cyber terror threat is growing, says Reid_Telegraph	SP		63	Report Chinese Hackers Stole Dalai Lamas Emails	PC
12	Apache got hacked	E		64	Report Hackers Steal South Korean Defense Documents	P
13	Cyber attack Threat on Rise Executives Say	SE		65	Spy Chief Says Cyber Attacks Work Of North Korea	PC
14	Cyber attacks Traced to N Korea	P		66	The Website of German Interior Minister Hacked	SP
15	Cyber Warfare On The Horizon	SP		67	China Gateway for Most Cyber-Attacks	P
16	Defaced govt websites another black eye for RP	P		68	Three alleged hackers arrested in Spain	S
17	Do Economic Crisis Administration Change Create Perfect Storm For Terror	SP		69	Three Spaniards arrested in alleged global hacking scheme	S
18	Europe vulnerable to cyber attack_guardian	SP		70	Tight security ahead of celebrations	PC
19	Experts dis rumored cyber-jihad set for Nov 11	SPC		71	Top websites targeted by hackers, experts warn	SE
20	_FBI to investigate Placentia Library hacking	SP		72	Twitter attack aimed at blogger	SP

21	FBI warns of cyber attack threat	SPC		73	US Electrical Grid Cyber spy Threat Could Mean Security Opportunity	P
22	FDIC Hackers took more than 120M in three months	SE		74	US_South Korea Cyber Attack Lessons Learned	S
23	Are we vulnerable to cyber-attacks	SE		75	Video hits websites eight days after anniversary	PC
24	Feds indict international cyber crook accused of 17Million ATM spree	E		76	Website defacing A new trend in hacking	P
25	Financial Crisis Called Top Security Threat to US	PE		77	With Unrest In Iran Cyber Attacks Begin	P
26	Google Finds Cyber Attacks on Vietnam Mine Dissidents	SPE		78	China hits back at Google's uncensored Hong Kong servers	SP
27	Hacker hits Iowa Homeland Security site and 2 others	PE		79	China Plays Constructive Role in tiding over global financial crisis	E
28	hacker holy wars	SPC		80	Chinese hackers steal secret Indian documents	P
29	Hackers Breach Iowa Gaming Commission Database	SPE		**The last 20 events are used to test for irrelevance to SPEC**		
30	Hackers infiltrate TD Ameritrade client database	E		1	As Autism Becomes More Common Docs Say Check Early	N/A
31	Hackers Run Wild and Free on AOL_wired	S		2	Bobby Petrino returning to Arkansas practice after motorcycle crash	N/A
32	Hackers steal FTP passwords of Symantec McAfee and others	E		3	Daniel Craig Talks Sky fall, Vows to Play James Bond Until They Tell Me to Stop	N/A
33	Hackers steal U.S.-S. Korean secrets	SPC		4	Dick Cheney leaves hospital after heart transplant	N/A
34	Brazils Blackout Spurs Hacker Speculation Alexander Mari-nis	P		5	Eagles say LT Peters has successful surgery to fix Achilles	N/A
35	Hackers Take Down the Most Wired Country in Europe	SC		6	FDA rejects call to ban BPA from food packaging	N/A
36	Hackers want to sabotage law enforcement agencies	S		7	Heat shoot for 16th straight home win face Sixers	N/A
37	High street chains next target for cyber terrorism	SC		8	Jim Carey, Jeff Daniels get Dumb and Dumber	N/A
38	Home Users Face Growing Risk of Cyber Attack	SE		9	Melanoma cases rising; young women at greatest risk	N/A
39	Independence Day Cyber Assault Whose Responsible	P		10	More evidence long-term estrogen therapy raises breast cancer risk	N/A
40	India has scary nuke hack	S		11	Red Sax closer Bailey may need thumb surgery	N/A
41	Inside the Chinese Hack Attack	P		12	Researchers warn kids don't get enough outdoor playtime	N/A
42	Internet security experts say Aurora attacks are very critical	S		13	RG3 turning down Colts workout is much ado about nothing; mail	N/A
43	Iranian Cyber Army hack of Twitter signals cyber-politics era	P		14	Sarah Palin on Today show Any GOP candidate would be better than Obama	N/A
44	Juniper Symantec investigating after Google attack	SE		15	Tigers place Inge on DL	N/A
45	Can India survive a Chinese cyber attack	P		16	The Beatles - The Next Generation' Paul McCartney's son says Fab Four's kids mulling idea	N/A
46	Law firm in Green Dam suit targeted with cyber attack	SPEC		17	'Titanic' director tweaks the sky	N/A
47	Lawmakers Electric utilities ignore cyber warnings	E		18	UC Riverside family autism resource center organizes effort to light up the bell tower blue to raise awareness of autism	N/A
48	Major US lab reports sophisticated cyber attack netted personal data on visitors	S		19	Venus wins Family Circle Cup opener	N/A
49	McAfee Inc Warns of Countries Arming for Cyber warfare	P		20	Wildcats live up to expectations with title	N/A
50	McAfee warns of new age cyber war	PE				

Table 5-5 Classification Result of Test 3

# Events	Expert	Classifier	Match Type	Match		# Events	Expert	Classifier	Match Type	Match
1	P	Null	Orientation	0.5		51	E	Null	Orientation	0.5
2	SP	Null	Not Match	N/A		52	PE	SP	Intersection	0.5
3	S	S	Exact	1		53	PE	Null	Orientation	0.5
4	SP	Null	Not Match	N/A		54	P	SP	Intersection	0.75
5	P	Null	Orientation	0.5		55	SP	Null	Not Match	N/A
6	SP	P	Membership	0.75		56	PC	P	Membership	0.75
7	E	Null	Orientation	0.5		57	SPE	Null	Not Match	N/A
8	SC	Null	Not Match	N/A		58	P	Null	Orientation	0.5
9	PS	Null	Not Match	N/A		59	P	Null	Orientation	0.5
10	SPE	SP	Membership	0.75		60	SE	Null	Orientation	0.5
11	SP	Null	Orientation	0.5		61	P	Null	Not Match	N/A
12	E	Null	Orientation	0.5		62	PE	E	Membership	0.75

13	SE	Null	Not Match	N/A		63	PC	SP	Intersection	0.5
14	P	P	Exact	1		64	P	SP	Intersection	0.75
15	SP	Null	Orientation	0.5		65	PC	P	Membership	0.75
16	P	SP	Intersection	0.75		66	SP	Null	Orientation	0.5
17	SP	Null	Orientation	0.5		67	P	SP	Intersection	0.75
18	SP	Null	Orientation	0.5		68	S	Null	Not Match	N/A
19	SPC	Null	Not Match	N/A		69	S	Null	Not Match	N/A
20	SP	P	Membership	0.5		70	PC	Null	Orientation	0.5
21	SPC	Null	Not Match	N/A		71	SE	SE	Exact	1
22	SE	E	Membership	0.75		72	SP	Null	Not Match	N/A
23	SE	Null	Not Match	N/A		73	P	Null	Not Match	N/A
24	E	Null	Orientation	0.5		74	S	Null	Not Match	N/A
25	PE	Null	Not Match	N/A		75	PC	Null	Orientation	0.5
26	SPE	P	Membership	0.5		76	P	Null	Orientation	0.5
27	PE	Null	Not Match	N/A		77	P	S	Not Match	N/A
28	SPC	Null	Orientation	0.5		78	SP	SP	Exact	1
29	SPE	Null	Orientation	0.5		79	E	Null	Not Match	N/A
30	E	E	Exact	1		80	P	P	Exact	1
31	S	Null	Not Match	N/A		81	Not	Null	N/A	N/A
32	E	Null	Orientation	0.5		82	Not	Null	N/A	N/A
33	SPC	P	Match	0.5		83	Not	Null	N/A	N/A
34	P	Null	Orientation	0.5		84	Not	Null	N/A	N/A
35	SC	Null	Not Match	N/A		85	Not	Null	N/A	N/A
36	S	Null	Not Match	N/A		86	Not	Null	N/A	N/A
37	SC	Null	Not Match	N/A		87	Not	Null	N/A	N/A
38	SE	Null	Orientation	0.5		88	Not	Null	N/A	N/A
39	P	P	Match	1		89	Not	Null	N/A	N/A
40	S	Null	Not Match	N/A		90	Not	Null	N/A	N/A
41	P	Null	Orientation	0.5		91	Not	Null	N/A	N/A
42	S	Null	Not Match	N/A		92	Not	Null	N/A	N/A
43	P	P	Exact	1		93	Not	Null	N/A	N/A
44	SE	SP	Intersection	0.5		94	Not	Null	N/A	N/A
45	P	P	Exact	1		95	Not	Null	N/A	N/A
46	SPEC	SP	Membership	0.5		96	Not	Null	N/A	N/A
47	E	Null	Orientation	0.5		97	Not	Null	N/A	N/A
48	S	Null	Not Match	N/A		98	Not	Null	N/A	N/A
49	P	Null	Orientation	0.5		99	Not	Null	N/A	N/A
50	PE	Null	Not Match	N/A		100	Not	Null	N/A	N/A

6 The User Interface of the Classifier

We developed a database-based web application to serve as a user interface of the classifier built with the approach proposed in this research.

6.1 Database Design

A relational database was first built which consists of three tables respectively for the text corpus, the key-word list and the stop words.

6.2 Web application

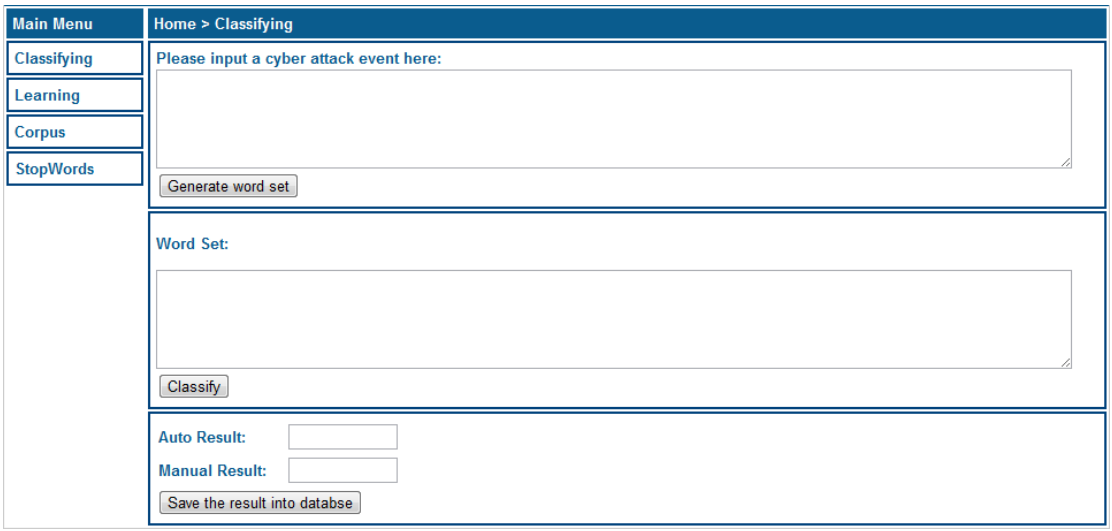
A web application was developed to implement the classifier. The Web application consists of four modules:

1) Classifying
2) Learning
3) Corpus
4) Stop words

6.2.1 Classifying

The user interface of this module is shown in Figure 6-1.

Figure 6-1 Classifying Page



The “Classifying” menu item allows the user to:

1) Input a text corpus to be classified.

2) Obtain a word set generated from the text input.

3) Obtain the automated classification result.

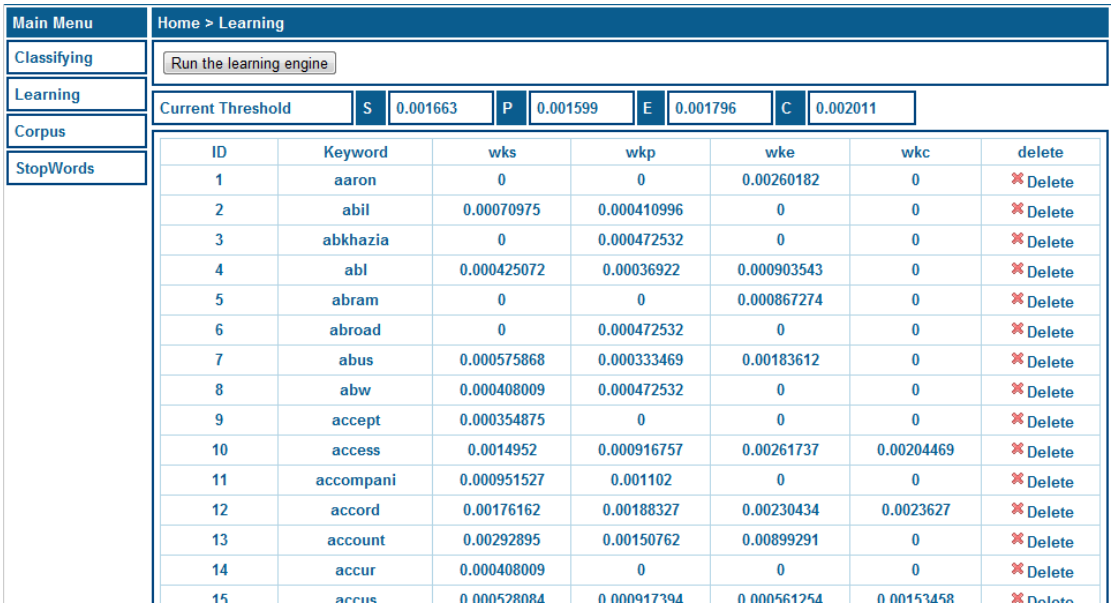
1) Provide their own judgment regarding the classification of the given text.

2) Have both of the automated classification result as well as the human decision saved in the corpus table of the database.

6.2.2 Learning

The user interface of this module is shown in Figure 6-2

Figure 6-2 Learning Page



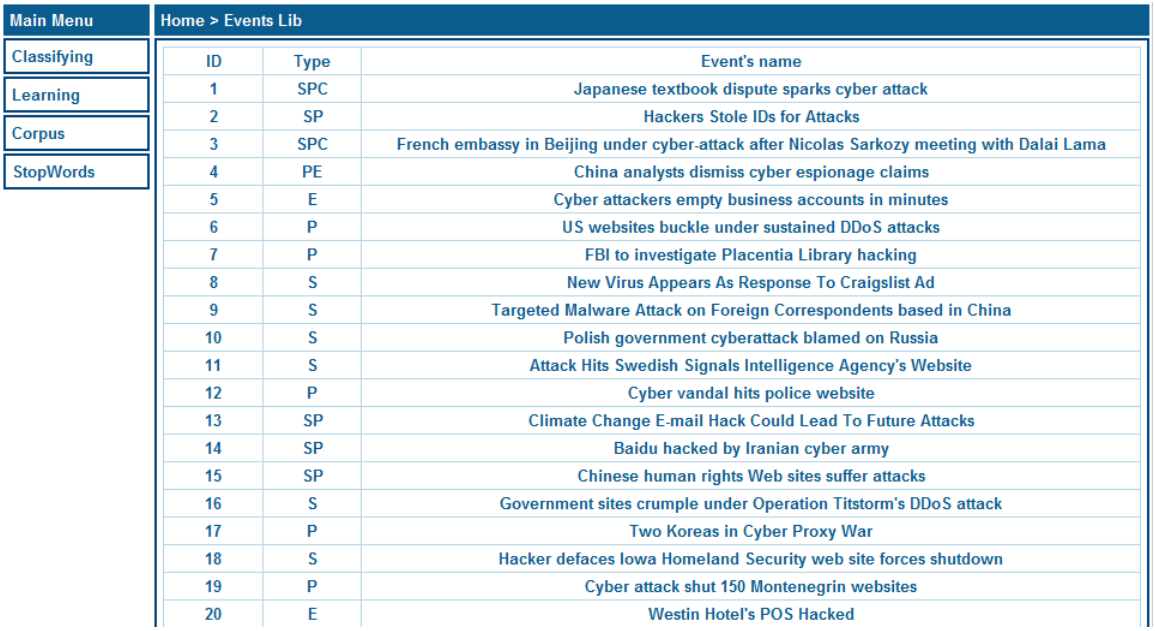
The “Learning” menu shows the key-word list as well as the thresholds with regard to the four categories S, P, E, and C. It is possible that, before a user enters this module, he has input a new event into the classifier, got a word set and the auto classification result generated, and has input his or her own judgment in the module of “Classifying”. As he or she does so, the corpus in the database is updated, but the key-word list shown here does not get updated automatically until the user clicks the

"Run the learning engine" button. The user can also manually delete individual keywords. This is useful when the user finds that some of the keywords have an extremely low term weight with respect to the four categories. By deleting some "useless" keywords, the running speed of the learning machine can be improved.

6.2.3 Corpus

The user interface of this module is shown in Figure 6-3.

Figure 6-3 Corpus Page

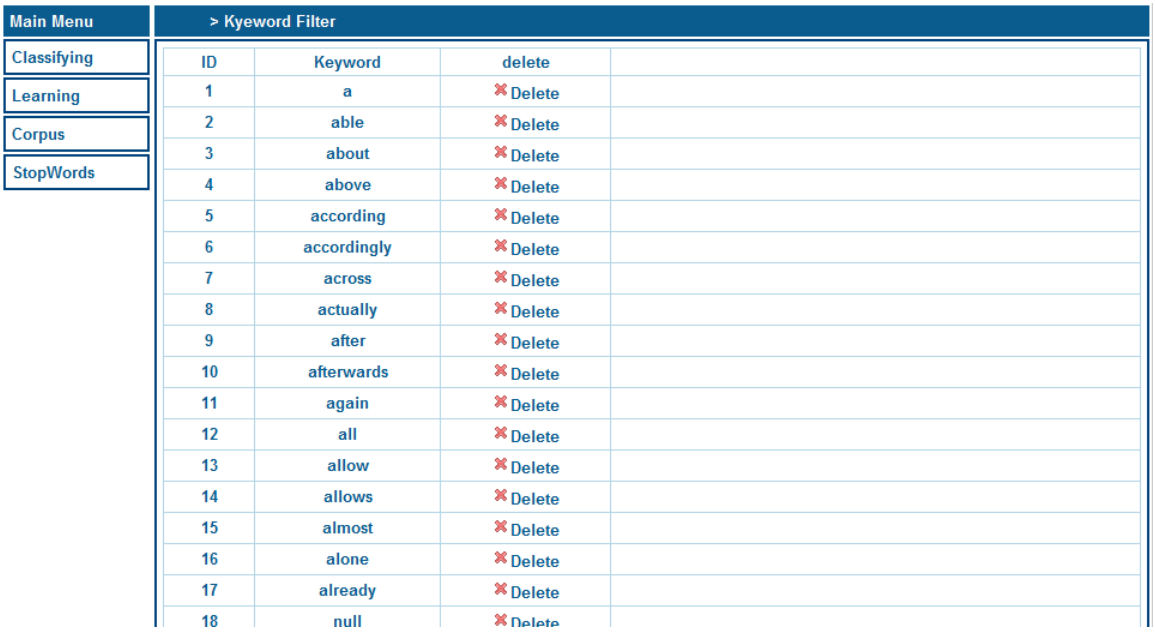


The "Corpus" menu serves as a visual, simplified text corpus list with cyber-attack events' names as well as their categories given by the expert decision displayed for users' reference.

6.2.4 Stop Words

The user interface of this module is shown in Figure 6-4.

Figure 6-4 Stop Words Page



The stop words used in the classifier construction are listed in this module. The default list is the one sorted by MIT which includes a total of 571 individual words [7]. The module allows users either to delete or add stop words at their choice.

6.3 Development Tools and Environment

The Database Management System used in this research is MySQL5.1; Apache Tomcat6 is used for client server communication; the web application was written in Java while the user interface was implemented in JSP and HTML.

7 Conclusion and Future Work

This thesis presents a development and application of a text categorization program in the context of cyber security. It was based on a previous study in which a collection of past cyber-attack events were classified into the Social, Political, Economic and Cultural categories. With the help of this taxonomy, it is hoped that the cyber-attacks can be better understood and/or predicted. This research was aimed at discovering a machine learning method which can realize a text classification task with respect to those cyber-attack categorization with acceptable effectiveness as well as efficiency. The method was based on the TF-IDF computation [17], but extended with several innovations.

7.1 Major innovations of the research

1) The research adopted a two-dimensioned table called the “key-word list” as a representation of machine learning results. The table contains the document indexing terms (i.e. key-words) and records the trained weights regarding their classification effects to each of the four cyber-attack categories. The weight was calculated by means of the TF-IDF algorithm.
2) A document (i.e. cyber-attack event) is classified by first searching in a keyword list, in which each word is stop-word-filtered and stem-reduced, for corresponding word matches. The weights of these matching words are accumulated for each of the S, P, E, and C category. The results were four variable values, named “the resemblance (with regard to this category),” each

representing the possibility of the document belonging to the particular category.

1) Adaptive threshold with regard to each of the four categories is developed. The threshold helps to determine the resemblance of a document with regards to a certain category. It is calculated in the learning process by a) comparing the resemblance values of all the documents regarding to one particular category, b) selecting the document with the smallest resemblance value from the *positive examples* and the document with the largest resemblance value from the *negative examples*, c) calculating the arithmetic mean of the two. The former is supposed to be larger than the latter, but if it is not true in some cases where the classification error is considerably influential, we try the second smallest and the second largest instead and so forth.

2) The classifier built in this research works efficiently in the sense that it could assign multiple labels to one document. While each document is classified with respect to the four categories, it is not necessary that a document must be uniquely assigned to one category. For example, it is possible that one document reaches or surpasses the threshold values of both the Political category and Social category in the classification process. In this case the document is classified in “P” and “S” simultaneously. Since one document might be attributed to more than one category, it makes the evaluation of the effectiveness technically a little bit more complicated than usual. To tackle this problem, a six-scaled criterion system was created to examine the degree to

which the text classification result matches the expert decision and hence if this text classification result can be justified. The calculation method of precision was adjusted accordingly so that the overall effectiveness of the classifier regarding to all of the four categories can be determined in one-step.

7.2 Directions of future work

The future work of this research can be taken in the following few directions:

1) The sample size in this study was relatively small. In the future, a primary task is to collect more cyber-attack records to enlarge our initial corpus and pre-classify them manually.
2) Instead of assigning a fixed value of 0.5 to an event that achieves an Orientation Match as its matching degree, we will calculate the exact value of this matching degree by applying an algorithm and thus further refine the classifier in the future.
3) A third research area is to substitute individual words for phrases as indexing terms, but use the same classifier-construction and evaluation approaches in order to gauge the effect on the classification effectiveness.
4) The current classification module of the web application can only process new-coming cyber-attack events in the form of text. The next goal is that it should allow the users to provide a URL, and have the application obtain the text corpus from the web site.
5) It is possible to further divide each category into several sub-categories. For example, the Social category might be divided into Education, Medicine, and

so on, which helps to further refine the features of different kinds of cyber-attacks.

This research, while successful, has demonstrated that currently utilized methods described in the literature can be improved upon. Future work as described here may make the classification effectiveness and ease of use even better.

8 Reference

[1] Antonie, M.-L., Zaïane, O. R., "Text Document Categorization by Term Association", http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1183881&tag=1

[2] Apté, C., Damerau, F.J., and Weiss, S.M. 1994, "Automated learning of decision rules for text categorization" *in ACM Transactions on Information Systems* 12, 3, pp. 333-251.

[3] Cybersecurity, retrieved from *Department of Homeland Security official website*, http://www.dhs.gov/cyber

[4] Dumais, S. T., Platt, J., Heckerman, D., and Sahami, M. 1998, "Inductive learning algorithms and representations for text categorization", in *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management* (Bethesda, US, 1998), pp. 148-155.

[5] Eom, J., Han, Y., *et al*. "Active Cyber-attack model for network system's vulnerability assessment", in *Proc. Int. Conf. Information Science and Security*, 2008, pp. 153-158.

[6] Gabrilovich, E and Markovitch, S., "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge", in *Proceedings of the National Conference on Artificial Intelligence* 21 (2), 1301

[7] Gandhi, R., Sharma, A., *el at*, "Dimensions of Cyber-Attacks Social, Political, Economic and Cultural" IEEE, 2011.

[8] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[9] http://tartarus.org/~martin/PorterStemmer/

[10] Kuhl, M. E., Kistner, J. *et al.* "Cyberattack modeling and simulation for network security analysis," in *Proc. Winter Simulation Conf.,* 2007, pp. 1180 –1188.

[11] Liu, Z., Wang, C. and Chen, S.; "Correlating multi-step attack and constructing attack scenarios based on attack pattern modeling," in *Proc. Int. Conf. Information Security and Assurance*, 2008, pp. 214–219.

[12] Markoff, J., "Internet attacks seen as more potent and complex," 2008,

http://www.iht.com/articles /2008/11/10/ technology/10attacks.php.
[13] Mitchell, T. M. 1996. Machine learning. McGraw Hill, New York, US.
[14] Myers, M. and Tan, F., "Beyond models of national culture in information systems research," *Advanced Topics in Global Information Management*, ch. 1, 2003.
[15] Peng, X. and Zhao, H., "A framework of attacker centric cyber-attack behavior analysis," in *Proc. IEEE Int. Conf. on Communications*, 2007, pp. 1449 –1454.
[16] Rasche, G., Allwein, E. *et al*. "Modelbased cyber security," in *Proc. 14th Annual IEEE Int. Conf. and Workshops on the Engineering of Computer-Based Systems*, 2007, pp. 405–412.
[17] Sebastiani, F. 2002. "Machine learning in automated text categorization" *ACM Computing Surveys* 34(1):1–47.
[18] Slay, J., "IS security, trust and culture: a theoretical framework for managing IS security in multicultural settings," *J. Campus-Wide Information Systems*, 2003, vol. 20, no. 3, pp. 98-104.
[19] Strategypage.com, "Information Warfare Article Index: Cyber War as the Ultimate Weapon," Jan. 5, 2008, http://www.strategypage.com/htmw/htiw/articles/20080105.aspx.